Дисциплина
«Методы автоматической обработки текстов»

Автоматическая обработка текста – преобразование текста на искусственном или естественном языке с помощью компьютера.

Text 1

Text 2

Сферы применения систем автоматической обработки текстов

Системы ***автоматической обработки текста*** (т.е. переработки одного вида текста в памяти ЭВМ в другой) по выполняемым функциям (входной и выходной информации) можно классифицировать следующим образом:

Язык входного текста	Язык выходного текста
ЕЯ-1	ЕЯ-2
Искусственный	Естественный
Естественный	Искусственный/Естественный
Естественный	Естественный (+Искусственный)

К системам первого типа относятся программы машинного перевода, получающие текст на некотором естественном языке и перерабатывающие его в текст на другом естественном языке.

Pragma

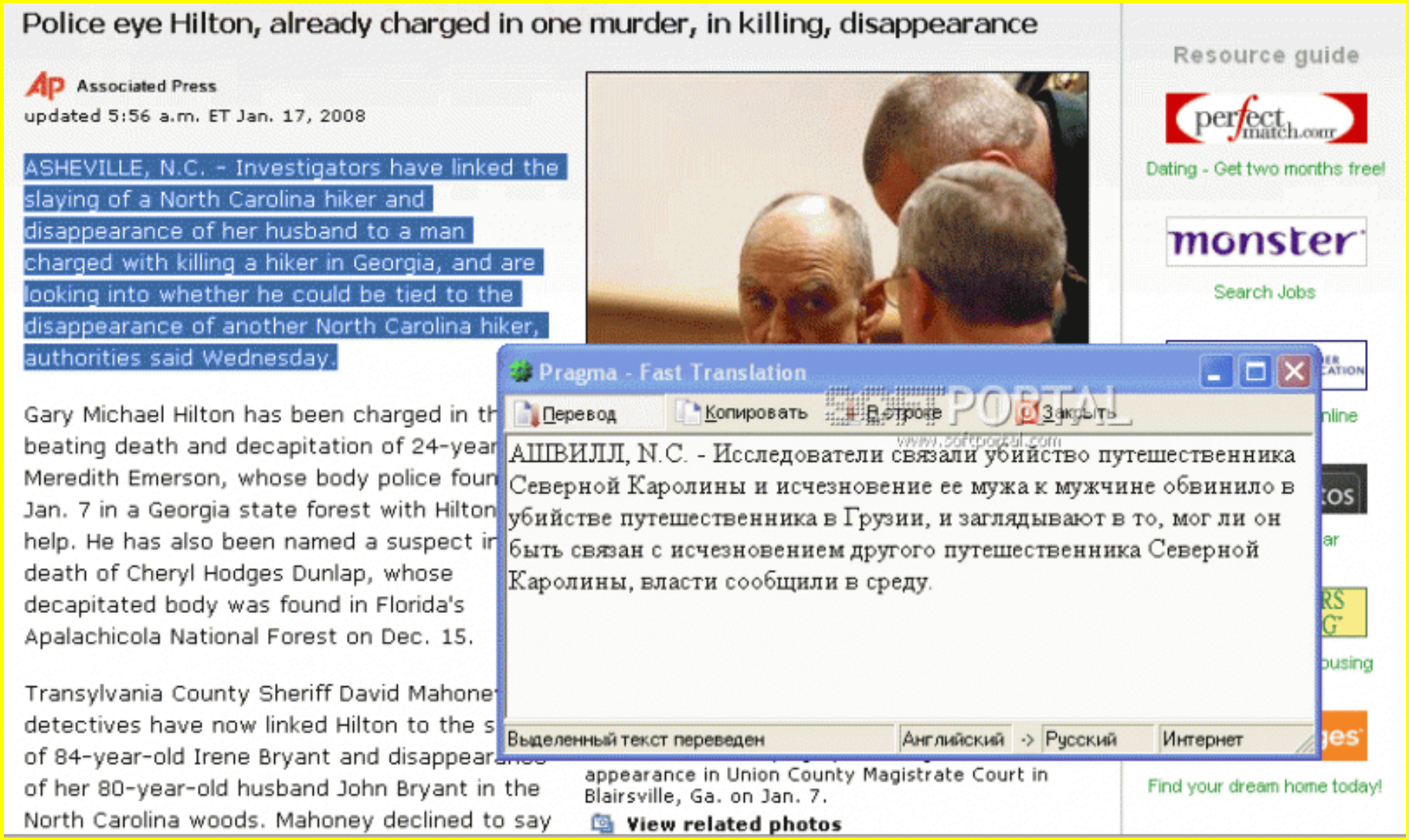
Promt

Lingvo

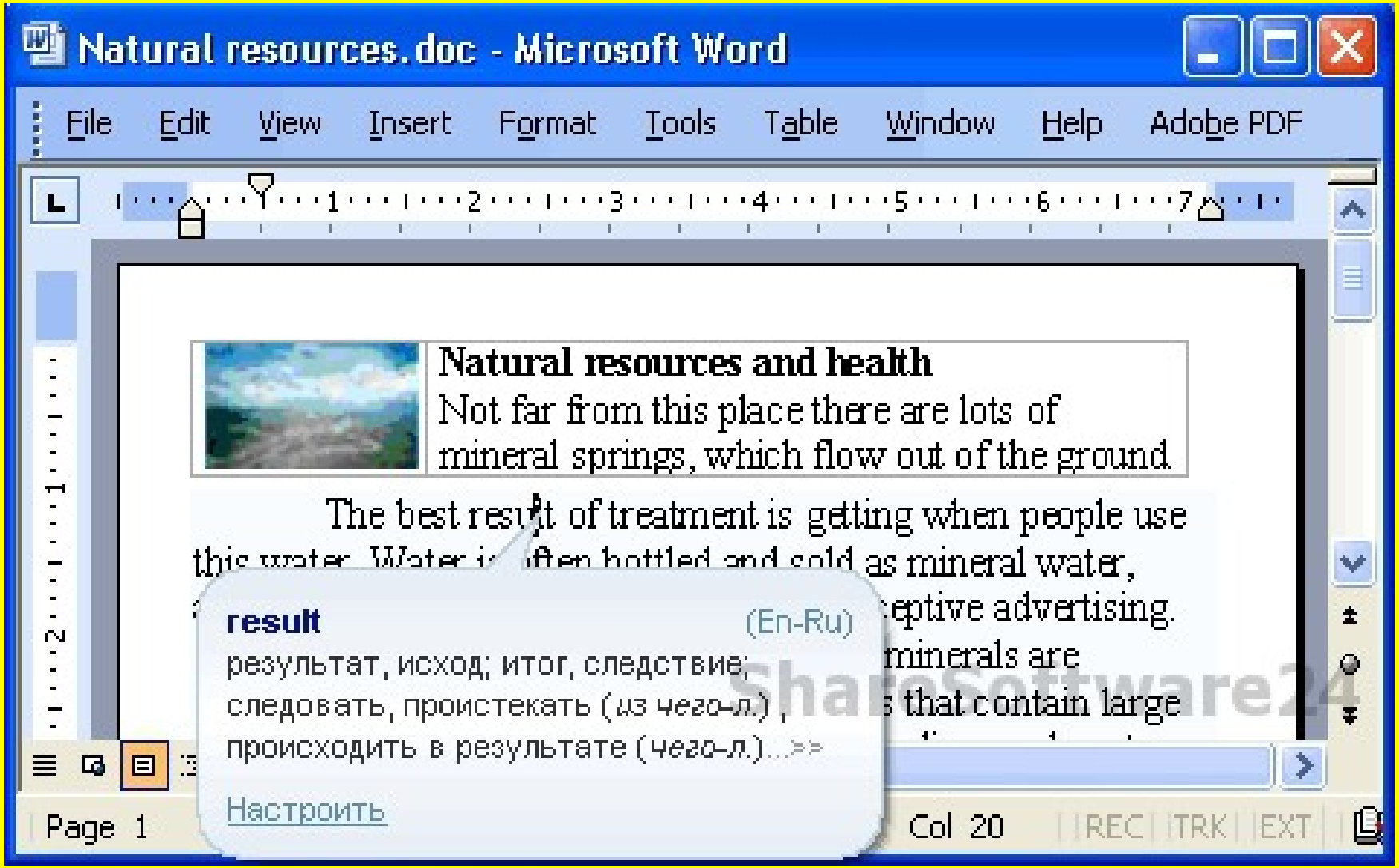
Socrat

Белазар

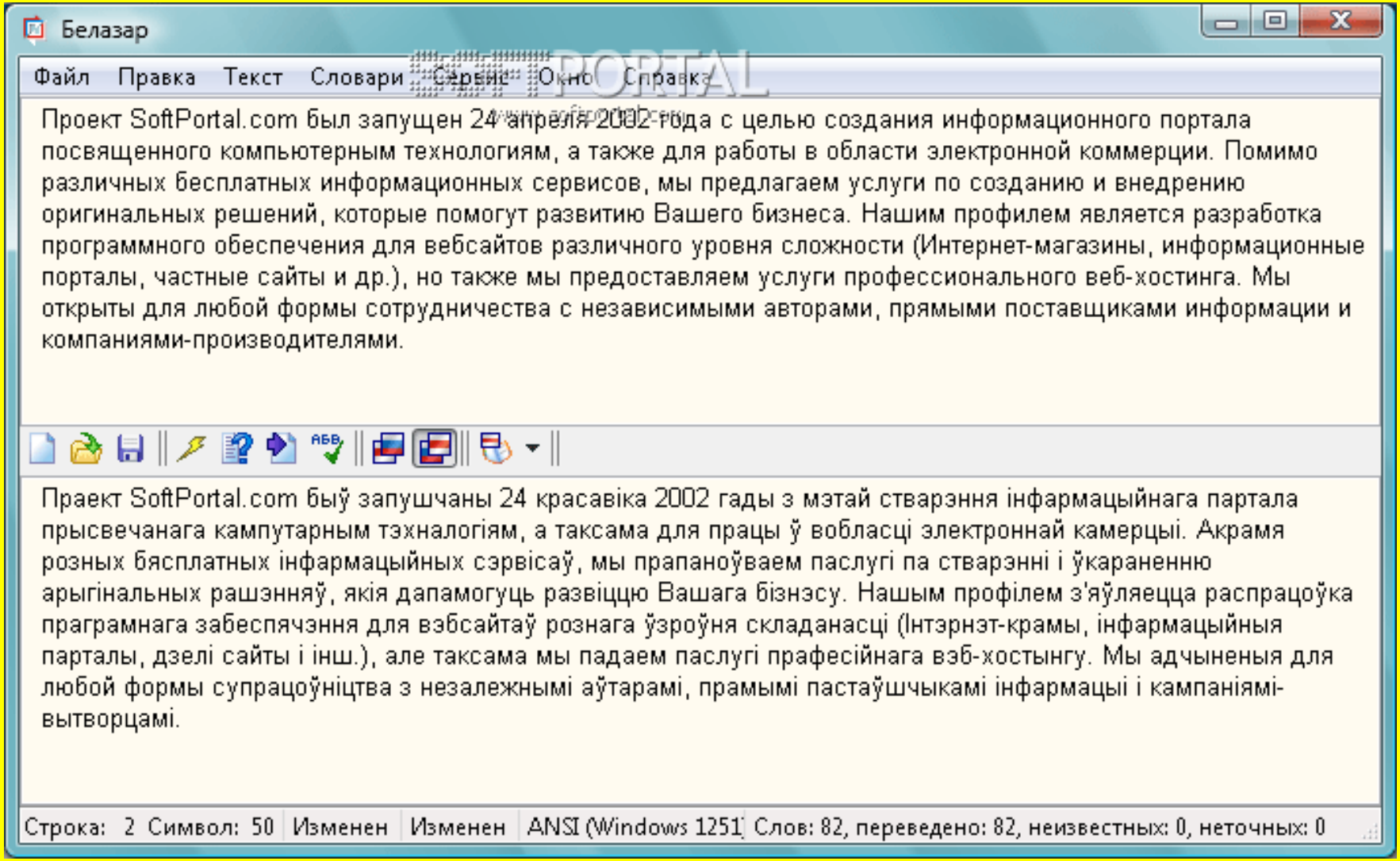
Online-перевод (например, Google Translator)



Скриншот программы Pragma



Lingvo 12, интегрированный в Microsoft Word



Окно переводчика Белазар, версия 6.1

Без Имени (Оригинал) - PROMT Editor Lite

Перевод | Настройка перевода | Правка | Сервис

Редактировать настройки
Сохранить как шаблон
Тематика документа

Словарная статья
Словарь

Незнакомые слова
Незнакомые слова

Зарезервировать слово
Резервирование

Редактировать
Шаблоны тематик

Project SoftPortal.com has been started on April, 24th, 2002 for the purpose of creation of an information portal devoted to computer technologies, and also for work in the field of electronic commerce. Besides various free information services, we offer services in creation and introduction of original decisions which will help development of your business. Our profile is working out of the software for

Тематика документа (Англо-Русский)

На основе шаблона "Общий"

Словари | Зарезервированные слова | Препроцессоры | Правила перевода | Другие настройки

Название	Значение
Выберите тип текста	тип не указан
Перевод местоимения "you"	Вы
Род местоимения "I"	мужской род
Род местоимения "you"	мужской род
Транслитерация имен собственных	не транслитерировать

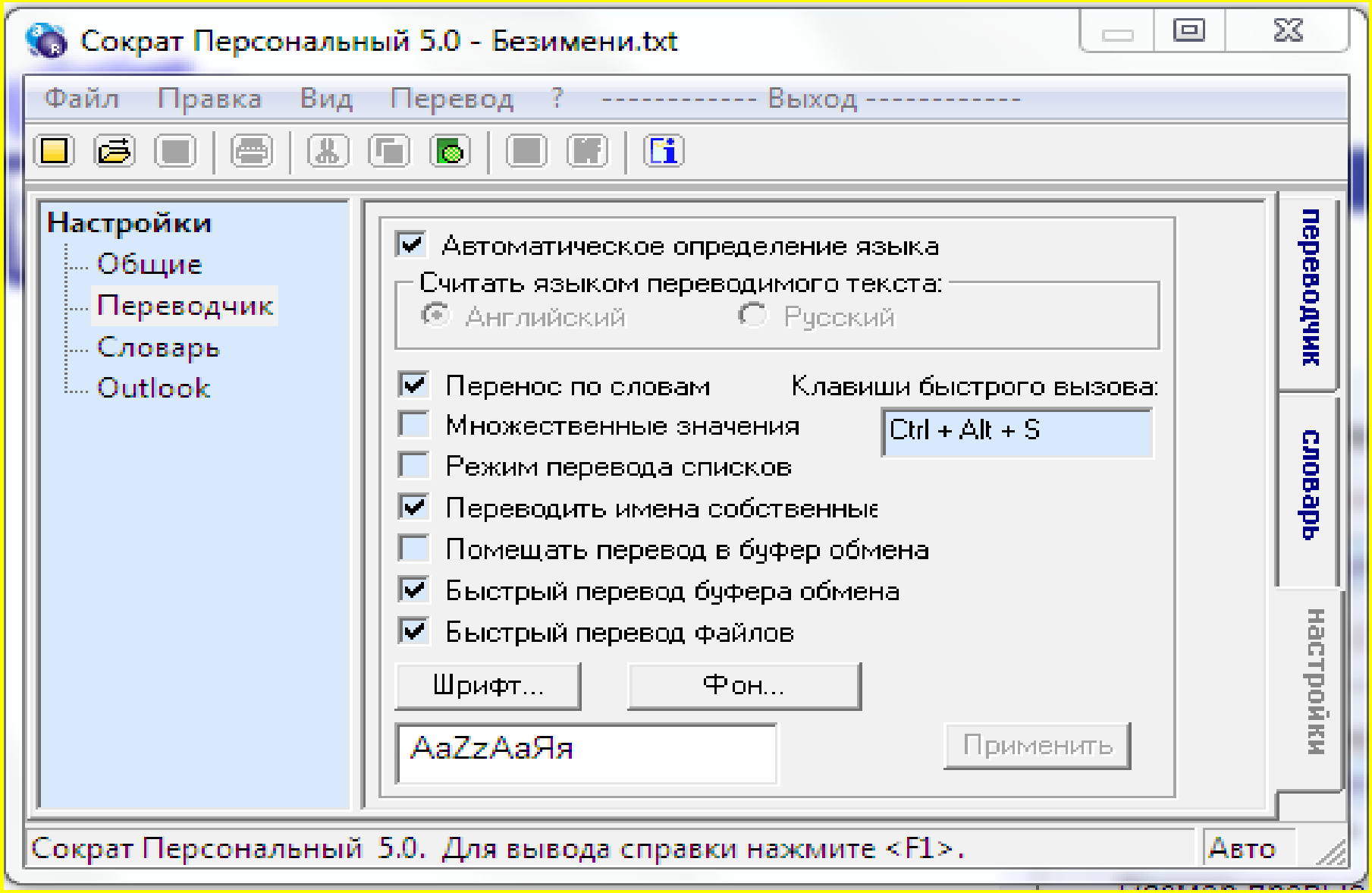
Выберите тип текста

Если вы выберете из списка один из указанных типов, то ряд конструкций будет переводиться на русский язык с учетом указанного вами стиля. Например, при выборе "документы, инструкции, контракты" предложение "When starting the program an error message appears" будет переводиться "При запуске программы появляется ошибочное сообщение". Его перевод без указания данного типа - "Запуская программу, появляется ошибочное сообщение". При выборе "рецепты" предложение "Place cracked wheat in bowl" переводится "Поместите раздробленную пшеницу в миску", без указания данного типа - "Место раздробило пшеницу в миске".

Сохранить как шаблон | Закрыть | Справка

Чтобы получить справку, нажмите F1 | Англо-Русский | Общий | CAP | NUM

Окно программы PROMT STANDARD 9.0 Giant



Окно программы Сократ Персональный 5.0

Второй тип - системы генерации (синтеза) текстов по некоторому формальному описанию.

Программы генерации текстов:

SciGen – программа для генерации псевдонаучных статей с графиками и списком литературы (на английском языке, сайт http://pdos.csail.mit.edu/scigen/).
Рифмач – программа для генерации поздравлений по заданным параметрам (пол, возраст, хобби, род занятий, имени характера, сайт http://rifmach.ru/)
Нонсенский генератор стихотворной продукции, фактически выдает набор строк и слов, объединенных в бессмысленное стихотворение с рифмой http://www.nonsence.de/generator/gen.php)
Textgen – платный генератор текстов на заданную тему, сайт http://www.textgen.ru/

Программы синтеза текстов (речи):
AT&T Labs Natural Voices
Linguatec Voice Reader
Imtranslator
Neospeech

SCIgen - An Automatic CS Paper Generator - Mozilla Firefox

Файл Правка Вид Журнал Закладки Инструменты Справка

Прикладное программное обеспеч... | Прикладное программное обеспеч... | CoGenTex, Inc. - Solutions | SCIgen - An Automatic CS Paper G

Назад Вперёд DownloadHelper http://pdos.csail.mit.edu/scige Обновить Остановить mail.ru: По Домой

@mail.ru socrat screenshot Найти! Почта 24 Мой мир 0,00 р. $ 30.53

Generate a Random Paper

Want to generate a random CS paper of your own? Type in some optional author names below, and click "Generate".

Author 1:
Author 2:
Author 3:
Author 4:
Author 5:

Generate Сброс

SCIgen currently supports Latin-1 characters, but not the full Unicode character set.

Examples

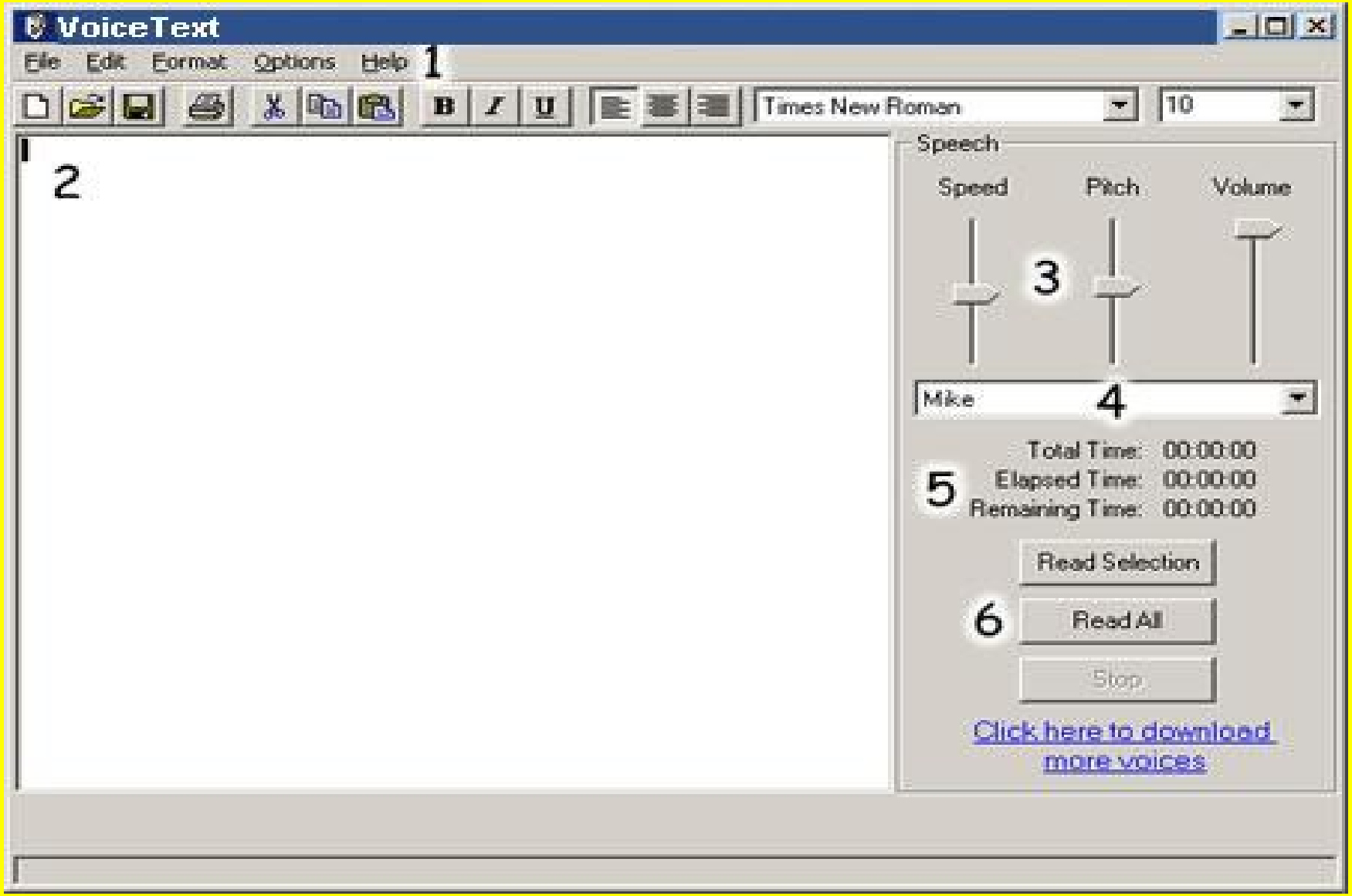
Here are two papers we submitted to WMSCI 2005:

- Rooter: A Methodology for the Typical Unification of Access Points and Redundancy (PS, PDF)
  Jeremy Stribling, Daniel Aguayo and Maxwell Krohn

  This paper was accepted as a "non-reviewed" paper!
  - Acceptance e-mail
  - A strange follow-up email, along with our response
  - Anthony Liekens sent an inquiry to WMSCI about this situation, and received this response, with an amazing letter (PS, PDF) attached. (Also check out Jeff Erickson's in-depth deconstruction of this letter.)
  - With the many generous donations we received, we paid one conference registration

Найти: нотир Следующее Предыдущее Подсветить все Учесть регистр

Сайт генератора текстов SCIGen



Окно программы NeoSpeech

Системы третьего типа перерабатывают текст на естественном языке в текст на искусственном (индексирование, извлечение смыслового содержания) или в другой текст на естественном языке (реферирование).

Индексирование осуществляет любая поисковая система, от поиска в системе Windows до интернет-поисковиков.

Программы для автоматического аннотирования и реферирования текстов – Либретто, Inxight Summerizer, Prosum, Text Referent, Extractor, Text Analyst.

К четвертому классу относятся программы, занимающиеся проверкой текста, написанного на естественном языке. Они в результате своей работы либо исправляют входной текст автоматически, либо формируют некоторый протокол замечаний.

Примеры – Microsoft Office Word, ABBYY Fine Reader, ORFO.

!!! Для создания систем, работающих со всем естественным языком без потери глубины анализа, в настоящий момент не хватает :

- либо технических возможностей (быстродействия, памяти),

- либо теоретической базы (например, пока нет даже единой схемы достаточно полного, глубокого и непротиворечивого описания семантики естественного языка).

Однако в коммерческих системах (предназначаются для большого количества пользователей) разных предметных областей, принята концепция поверхностного анализа, к тому же и производится такой анализ значительно быстрее.

Дальнейшее продвижение вперед, использование естественного языка в практических областях невозможно без оснащения этих систем ***обширными*** и ***глубокими*** (с точки зрения охвата различных явлений языка) описаниями и моделями, созданными лингвистами-профессионалами.

В настоящее время создаются АОТ-системы, представляющие коммерческий интерес и использующихся при решении следующих прикладных задач:

1.Machine Translation and Translation Aids - машинный перевод;

2.Text Generation - генерация текста;

3.Localization and Internationalization - локализация и интернационализация;

4.Controlled Language - работа на ограниченном языке;

5.Word Processing and Spelling Correction - создание текстовых документов (ввод, редактирование, исправление ошибок)

6.Information Retrieval - информационный поиск и связанные с ним задачи.

Машинный перевод

Исторически ***машинный перевод*** является первой попыткой использования компьютеров для решения невычислительных задач (Джорджтаунский эксперимент в США в 1954 г.; работы по машинному переводу в СССР, начавшиеся в 1954 г.).

Однако первоначальный оптимизм по поводу того, что столь трудоемкую работу можно поручить ЭВМ, сменился разочарованием в связи с абсолютной непригодностью получаемых текстов.

Robin, Robin, what a man!
He eats as much as no one can.
He ate a lot of fish, he ate a lot of meat.
He ate a lot of ice-cream and a sweet.
He ate a lot of porridge and ten eggs
And all the cookies Mother had.
He drank a lot of juice, he ate a cake
Then said: "I have a stomach-ache"

Робин Бобин Барабек
Скушал сорок человек.
И корову, и быка,
И кривого мясника,
И телегу, и дугу,
И метлу, и кочергу.
Скушал церковь, скушал дом,
И кузницу с кузнецом,
А потом и говорит:
– У меня живот болит!

Малиновка, Малиновка, какой человек!
Он ест насколько никто не может.
Он съел много рыб, он съел много мяс.
Он съел много ледяных-сливки и сладкий.
Он съел много каша и десять яйцо
И вся Мать повары имела.
Он пил много соков, он съел торт
Затем сказал: "У меня есть желудок-боль"

Результат перевода в одной из коммерческих переводческих систем

Квалифицированный переводчик понимает смысл текста и **пересказывает** его на другом языке словами и стилем, максимально близкими к оригиналу.

Для компьютера этот путь выливается в решение двух задач:

- перевод текста в некоторое внутреннее семантическое представление
- генерация по этому представлению текста на другом языке.

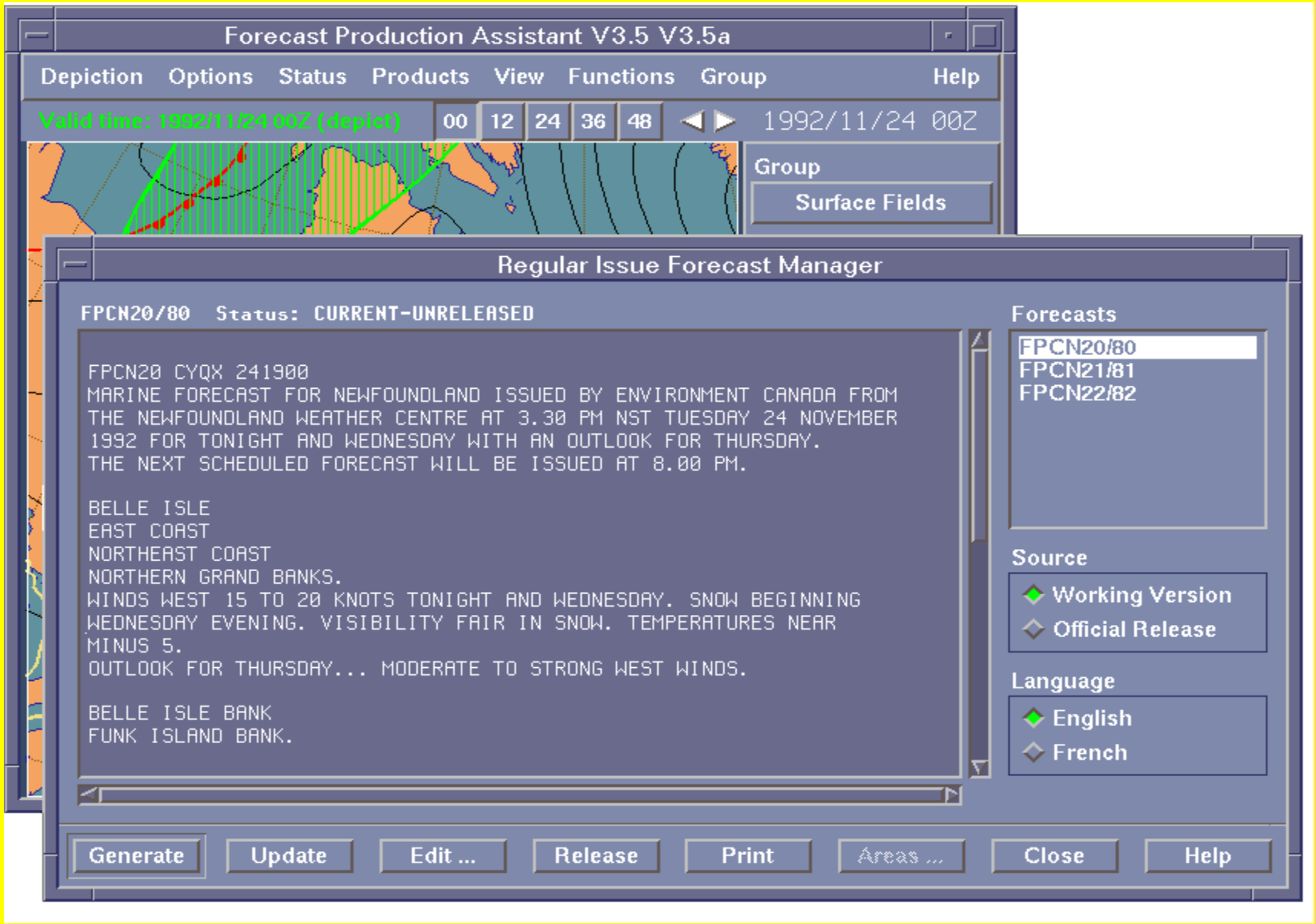
Более реалистичными являются попытки создать системы ***автоматизированного перевода*** - программы, которые не берут на себя полностью весь перевод, а лишь помогают человеку-переводчику справиться с некоторыми трудностями (Computer Aided Translation).

В последнее время также появляются автоматизированные системы **«доперевода»** или **«перевода изменений»**.

Их возникновение связано с тем, что большинство технических текстов (описания, инструкции) не являются целиком новыми (как и явления, продукты, механизмы и т.п., ими описываемые), а содержат в себе лишь некоторые изменения, связанные, например, с усовершенствованием конструкции.

Система «доперевода» извлекает из памяти знакомые предложения, а новые куски предлагает переводчику.

Развитием систем подобного вида можно считать канадскую систему генерации прогнозов погоды Forecast Generator (FOG). Можно считать, что в ней перевод полностью заменен генерацией текстов. В памяти системы хранится 20 миллионов слов и словосочетаний, связанных с прогнозами погоды, что позволяет генерировать как английский, так и французский вариант непосредственно из базы данных.



Окно системы ForeCast Generator

Генерация текста

По степени сложности и выразительности существующие методы генерации сообщений принято подразделять на 4 класса (часто используются комбинации методов).

1) ***Canned-based methods***
Неизменяющийся шаблон - просто печать строки символов без каких-либо изменений (1 file copied, а в случае, например, трех - 3 files copied)
2) ***Template-based methods***
Изменяющийся шаблон - бесконтекстная вставка слов в образец-строку (именно этот метод используется в MS-DOS): 0 file(s) copied, 1 file(s) copied, 2 file(s) copied).
3) *Phrase-based methods*
Контекстная вставка.
В зависимости от вида сообщения (контекста) шаблон может быть несколько изменен (1 file copied, 2 marked files copied, 2 marked files deleted).
4) *Feature-based methods*
Синтез сообщения на основе набора свойств (грамматических признаков): 1 file should be copied, 1 file was copied, 2 marked files were copied

Генерация логически связных, целостных текстов является более сложной задачей: к правилам построения предложений добавляются правила их сочетаемости, правила развития сюжета, соблюдения стиля и т.п.

Для ряда специальных текстов данные правила оговорены некоторыми стандартами, немногочисленны и поддаются формализации.

Примерами подобных текстов могут служить различные инструкции, техническая документация.

Поиски решения этих проблем привели в свое время к появлению новой профессии «технического писателя».

В последние годы появились практические системы, осуществляющие помощь в разработке документации, вплоть до ее автоматической генерации.

Форма и содержание документации часто выбирается не столько из соображений удобства и полезности для пользователя, сколько из соображений простоты ее создания.

Очевидно, что качественная система должна генерировать текст, **правильный** с точки зрения грамматики и синтаксиса естественного языка.

Ввиду четкого определения предметной области и строгости правил при составлении технических документов степень формализации в постановке данной задачи существенно выше, чем в задаче машинного перевода, что позволяет надеяться на более высокие результаты.

Генератор текста — компьютерная программа, способная генерировать последовательности символов, внешне похожие на текст, но при этом, как правило, лишённые смысла (в связи с чем генераторы текста иногда называют «генераторами бреда», «генераторами шизофазии»).

Тексты, созданные с помощью генераторов, являются правильными с точки зрения большинства языковых норм.

Файл Правка Вид Журнал Закладки Инструменты Справка

Прикладное программное обеспечение... | Генератор текста — Википедия | Яндекс.Рефераты

Назад Вперёд DownloadHelper http://vesna.yandex.ru/pushkin/ Обновить Остановить mail.ru: Пс Домой Firebu

@mail.ru генераторы текстов Найти! Почта 28 Мой мир 0,00 р. $ 30.53 € 42.2

Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё

Яндекс рефераты

Рефераты Пушкин Кризйтор Гадалка

Написать:

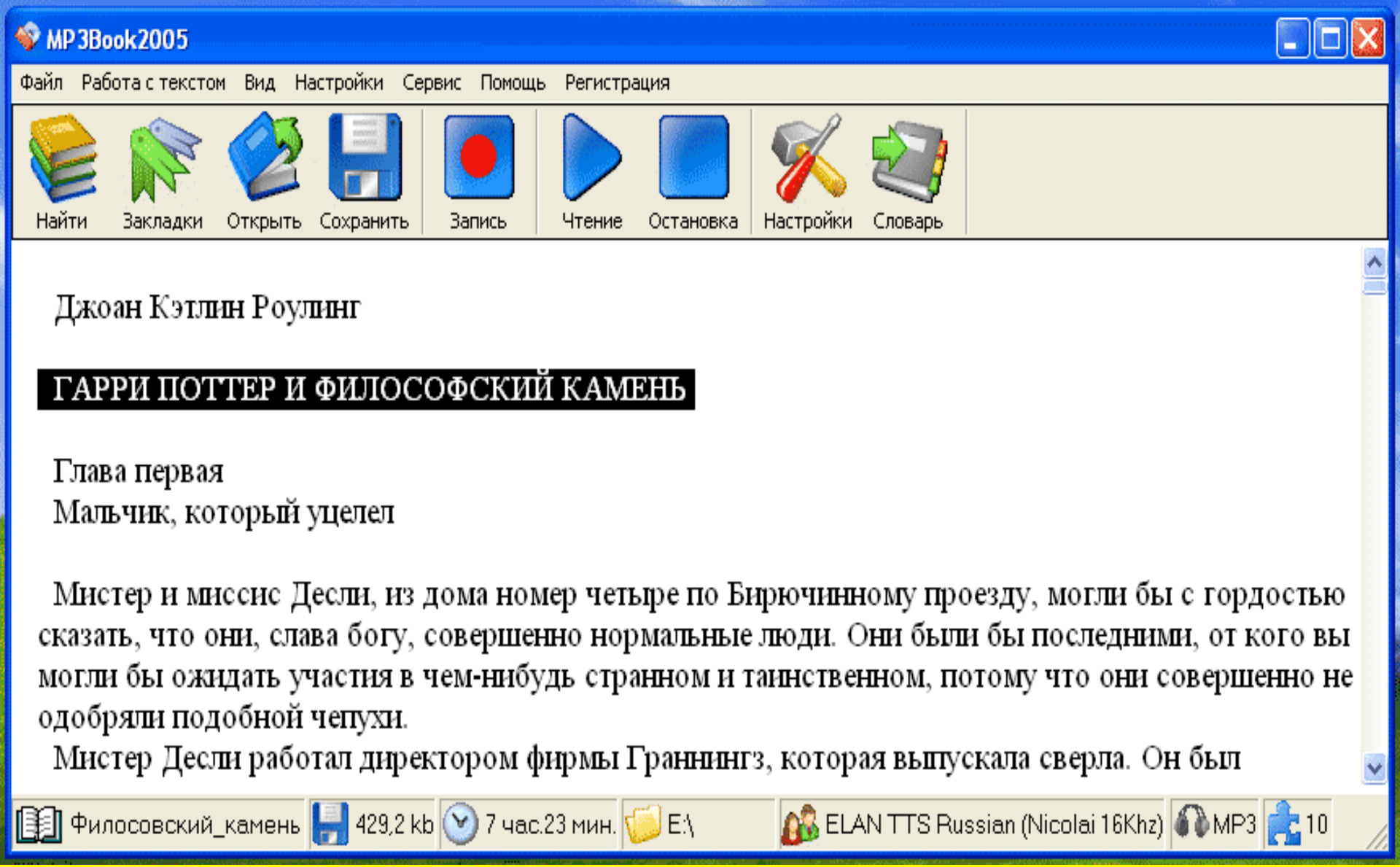
— онегинские строфы
— одические строфы
— октавы
— терцины
— сонет

6 июня — день рождения Александра Сергеевича Пушкина. К этому празднику мы подготовили новый генератор текстов. На этот раз — зарифмованных. Генератор стихов им. А.С.Пушкина поможет новичкам сделать первые шаги, а людям, хорошо знакомым с творчеством поэта — проверить себя, угадав, из каких произведений взяты те или иные строки.Вы спросите, почему генератор стихов носит имя Пушкина? По той же причине, что и музей изобразительных искусств, и детская библиотека, и пароход.

©1997—2011 «Яндекс»

Найти: urol Следующее Предыдущее Подсветить все Учесть регистр

Ресурс ВЕСНА от компании ЯНДЕКС



Главное окно программы для синтеза речи MP3Book2005